Spatio-Temporal Modelling and Forecasting of Fine Particulate Matter

Modellizzazione e previsione spazio-temporale delle polveri fini

Sujit K. Sahu
S^{3}RI, School of Mathematics, University of Southampton, SO17 1BJ, UK
Email: S.K.Sahu@maths.soton.ac.uk

Summary: Studies indicate that even short-term exposure to high concentrations of fine atmospheric particulate matter ($PM_{2.5}$) can lead to long-term health effects. Data are typically observed at fixed monitoring stations throughout a study region of interest at different time points. The study region may contain both rural and urban areas. Statistical spatio-temporal models are appropriate for modelling these data.

In this talk I will summarise my recent work on modelling and short-term forecasting of $PM_{2.5}$ levels. I will talk about a a random effects model developed in Sahu *et al.* (2004) and briefly mention a Bayesian Kriged-Kalman filtering model detailed in Sahu and Mardia (2005). In the first approach we introduce two random effects components, one for rural or background levels and the other as a supplement for urban areas. These are specified in the form of spatio-temporal processes. Weighting these processes through population density results in nonstationarity in space. In the talk I will analyze a dataset on observed $PM_{2.5}$ in three states in the U.S. - Illinois, Indiana and Ohio.

Keywords: Bayesian Kriged-Kalman filtering; forecasting; spatio-temporal processes; fine atmospheric particulate matter.

1. Introduction

Particulate matter (PM) has been linked to widespread public health effects, including a range of serious respiratory and cardiovascular problems, and to reduced visibility in many parts of the United States. Additionally, these particles and their components interact in ways that contribute to elevated concentrations of other air pollutants and stress to vegetation and ecosystems. In 1997, the U.S. Environmental Protection Agency (EPA) promulgated new regulations that established National Ambient Air Quality Standards (NAAQS) for particulate matter (PM) with aerodynamic diameters less than 2.5μm ($PM_{2.5}$). As part of the program to implement these standards, a network of ambient mass monitoring sites was established. In 2003, over 950 sites were in operation in the U.S. These sites are located primarily in populated regions, measuring pollution where people live and work. Most of these sites are used to evaluate compliance with particulate air quality standards, other sites are located away from urban areas to characterize transport, background concentrations, and visibility levels.

PM can be emitted directly or formed in the atmosphere. Generally, $PM_{2.5}$ contains particles formed in the atmosphere from gaseous emissions. Examples include sulfates formed from sulfur dioxide (SO_2) emissions, nitrates formed from NO_x emissions, and carbon formed from organic gas emissions. Regional and local PM concentrations are affected by emissions, topography, land cover, and a variety of processes affecting the rates of conversion of gases to particles. The general pattern of air movement across

the U.S. also influences particle levels. During summer, generally, a south to north or northeast transport direction is found over the eastern U.S.

In recent years, hierarchical Bayesian approaches for spatial prediction of air pollution have been developed, see e.g. Le *et al*. (1997), Cressie *et al*. (1999), Kibria *et al*. (2000) Zidek *et al*. (2002). Smith *et al*. (2003) proposed a spatio-temporal model for predicting weekly averages of $PM_{2.5}$ and other derived quantities such as annual averages within three southeastern states. The $PM_{2.5}$ field is represented as the sum of semi-parametric spatial and temporal trends, with a random component that is spatially correlated, but not temporally.

In this talk we present a hierarchical space-time model, developed in Sahu *et al.* (2004), that introduces two spatio-temporal processes, one capturing rural or background effects, the second adding extra variability for urban/suburban locations. We also consider the relationship of population density to fine particulate matter and incorporate non-stationary spatial and temporal covariance structure, see Section 2.Estimates of the probabilities of non-compliance with the proposed air quality standard for annual $PM_{2.5}$ are also provided, based on the weekly predictions of $PM_{2.5}$ for 2001.

Sahu and Mardia (2005) present a short-term forecasting analysis of $PM_{2.5}$ data in New York City during 2002. Within a Bayesian hierarchical structure, they model the spatial structure with principal Kriging functions and the time component is modeled by a vector random-walk process following Mardia *et al.* (1998). In Section 3we provide the forecast distributions and the cross-validation statistics proposed by Sahu and Mardia (2005).

2. Hierarchical models

The model developed here is applicable for spatio-temporal data recorded at n sites $\mathbf{s}_i\,(\in \mathbb{R}^2), i = 1, \ldots, n$, over a period of T equally spaced time points, $t_1, \ldots, t_T$. Let $z(\mathbf{s}_i, t)$ denote the square-root of the observed $PM_{2.5}$ level at site $\mathbf{s}_i$ and at time t where $i = 1, \ldots, n$ and $t = 1, \ldots, T$.

Let $\alpha(\mathbf{s}_i)$ denote the indicator of the urban sites, i.e., $\alpha(\mathbf{s}_i) = 1$ if the site $\mathbf{s}_i$ is an urban site, =0 otherwise. In the model we also include population densities defined by $p(\mathbf{s}) = \sqrt{p'(\mathbf{s})/\max p'(\mathbf{s})}$, where $p'(\mathbf{s})$ denotes the population density at site $\mathbf{s}$ and the maximization is performed over all the sampled and the predictive sites. To model the seasonal effects we define the monthly seasonal indicator, $u(t_k, m)$ as follows: $u(t_k, m) = 1$ if the time t_k is in the mth month and zero otherwise, for $m = 1, \ldots, 12$. Let

$$\mathbf{x}(\mathbf{s}_i, t_k) = \big(1, p(\mathbf{s}_i), \alpha(\mathbf{s}_i), \alpha(\mathbf{s}_i) \times p(\mathbf{s}_i), u(t_k, 2), \ldots, u(t_k, 12)\big)'$$

for $i = 1, \ldots, n$ denote the covariate values and seasonal dummies. Define the mean function $\mu(\mathbf{s}_i, t_k)$ given by

$$\mu(\mathbf{s}_i, t_k) = \beta_0 + \beta_1\, p(\mathbf{s}_i) + \beta_2\, \alpha(\mathbf{s}_i) + \beta_3\, \alpha(\mathbf{s}_i) \times p(\mathbf{s}_i) + \sum_{m=2}^{12} \gamma_m u(t_k, m). \qquad (1)$$

Since β_0 is included in (1), for identifiability purposes we do not include $u(t_k, 1)$ in the model. Thus the unknown parameters in $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \gamma_2, ..., \gamma_{12})'$ are used to describe the mean structure of the data.

The first step in our spatio-temporal model is to assume the hierarchical structure:

$$Z(\mathbf{s}_i, t_k) = Y(\mathbf{s}_i, t_k) + \epsilon(\mathbf{s}_i, t_k),\ i = 1, \ldots, n,\ \ k = 1, \ldots, T, \tag{2}$$

where $Y(\mathbf{s}, t)$ is a space-time process and the error term $\epsilon(\mathbf{s}_i, t_k)$ is a white noise process and specifically assumed to follow $N(0, \sigma^2_\epsilon)$ independently.

The space-time process $Y(\mathbf{s}, t)$ is expressed as

$$Y(\mathbf{s}_i, t_k) = \mu(\mathbf{s}_i, t_k) + w(\mathbf{s}_i, t_k) + p(\mathbf{s}_i)\, v(\mathbf{s}_i, t_k), \tag{3}$$

where $w(\mathbf{s}, t)$ and $v(\mathbf{s}, t)$ are independent zero mean spatio-temporal processes. For convenience, for each of the processes, we adopt a separable covariance structure. That is,

$$\begin{aligned} \text{Cov}\big\{w(\mathbf{s}_i, t_l),\ w(\mathbf{s}_j, t_k)\big\} &= \sigma^2_w\, \rho_{sw}(\mathbf{s}_i - \mathbf{s}_j; \phi_{sw})\, \rho_{tw}(t_l - t_k; \phi_{tw}), & (4)\\ \text{Cov}\big\{v(\mathbf{s}_i, t_l),\ v(\mathbf{s}_j, t_k)\big\} &= \sigma^2_v\, \rho_{sv}(\mathbf{s}_i - \mathbf{s}_j; \phi_{sv})\, \rho_{tv}(t_l - t_k; \phi_{tv}). & (5) \end{aligned}$$

In addition, the ρ's are taken to be exponential correlation functions, i.e., $\rho(d; \phi) = \exp(-\phi\, ||d||)$.

What is our motivation for introducing two random processes in (3)? Here $w(\mathbf{s}, t)$ is viewed as a rural or background zero mean process; $v(\mathbf{s}, t)$ adds urban/suburban spatio-temporal uncertainty. In other words, β_0 is the overall mean level associated with the rural sites with $\beta_0 + \beta_2$ being the overall mean level for the urban/suburban sites.

Denote the unknown parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2_w, \sigma^2_v, \sigma^2_\epsilon)'$. We assume that, a priori, the β's are independent with distribution $N(0, A^2)$. We take A^2 to be large for vague prior specification. For the three variance parameters σ^2_ϵ, σ^2_w and σ^2_v we assume independent inverse gamma prior distributions, $IG(a, b)$ (with mean $b/(a-1)$) setting $a = b = \delta$ for a small positive value of δ to have a proper but vague prior distribution for each. We choose the decay parameters $\boldsymbol{\phi} = (\phi_{sw}, \phi_{sv}, \phi_{tw}, \phi_{tv})'$ using a Bayesian model choice criterion.

3. Forecasts and cross-validation

The posterior predictive distributions are used to make step ahead predictions (forecasts). Let $\mathbf{Z}_t = (Z(\mathbf{s}_1, t), \ldots, Z(\mathbf{s}_n, t))'$ denote the vector of random observations at time time t and $\boldsymbol{\xi}$ denote all the unknowns: the parameters $\boldsymbol{\theta}$, the space time processes $Y(\mathbf{s}, t)$, $w(\mathbf{s}, t)$, and $v(\mathbf{s}, t)$.

The 1-step ahead forecast distribution is given by,

$$\pi(\mathbf{z}_{T+1} | \mathbf{z}_1, \ldots, \mathbf{z}_T) = \int \pi(\mathbf{z}_{T+1} | \boldsymbol{\xi})\, \pi(\boldsymbol{\xi} | \mathbf{z}_1, \ldots, \mathbf{z}_T)\, d\boldsymbol{\xi}, \tag{6}$$

where the likelihood term $\pi(\mathbf{z}_{T+1} | \boldsymbol{\xi})$ is obtained from the hierarchical model (2). The L-step ahead predictions where $L > 1$ is a positive integer are obtained similarly. The mean of the forecast distributions like (6) are the optimal forecasts under a squared error loss function.

For cross-validation purposes Sahu and Mardia (2005) developed a weighted distance between the forecasts and the actual observations. Let $\mathbf{U} = \big(\mathbf{Z}'_{T+1}, \ldots, \mathbf{Z}'_{T+L}\big)'$ denote the set of observations for which we seek validation. Note that we have observed data $\mathbf{Z}_1, \ldots, \mathbf{Z}_{T+L}$ but we have used only $\mathbf{Z}_1, \ldots, \mathbf{Z}_T$ to fit the model and obtain the validation forecast for $\mathbf{U}$. Let $\mathbf{u}_{\text{obs}}$ denote the observed data.

Using the implemented MCMC, we draw $\mathbf{U}^{(j)}, j = 1, \ldots, E$ (where E is a large positive integer) samples from the forecast distribution $\pi(\mathbf{u}|\mathbf{z}_1, \ldots, \mathbf{z}_T)$. Now

$$\bar{\mathbf{U}} = \frac{1}{E}\sum_{j=1}^{E}\mathbf{U}^{(j)}, \text{ and } \hat{\Sigma} = \frac{1}{E-1}\sum_{j=1}^{E}\left(\mathbf{U}^{(j)} - \bar{\mathbf{U}}\right)\left(\mathbf{U}^{(j)} - \bar{\mathbf{U}}\right)',$$

unbiasedly estimate the mean vector and the covariance matrix of the forecast distribution $\pi(\mathbf{u}|\mathbf{z}_1, \ldots, \mathbf{z}_T)$, respectively. Under suitable regularity conditions which guarantee asymptotic normality and for small values of L, the predictive distribution $\pi(\mathbf{u}|\mathbf{z}_1, \ldots, \mathbf{z}_T)$ can be approximated by the nL-dimensional normal distribution with mean $\bar{\mathbf{U}}$ and covariance matrix $\hat{\Sigma}$. Using well-known properties of multivariate normal distribution, we have,

$$D^2 = (\mathbf{U} - \bar{\mathbf{U}})'\hat{\Sigma}^{-1}(\mathbf{U} - \bar{\mathbf{U}}) \sim \chi^2_{nL}, \text{ approximately.} \tag{7}$$

The approximation arises due to the fact that $\mathbf{U}$ is only approximately multivariate normal for small values of L for short-term forecasting.

The validation statistics proposed by Sahu and Mardia (2005) is the observed value of D^2 given by,

$$D^2_{\text{obs}} = (\mathbf{u}_{\text{obs}} - \bar{\mathbf{U}})'\hat{\Sigma}^{-1}(\mathbf{u}_{\text{obs}} - \bar{\mathbf{U}}). \tag{8}$$

Clearly, D^2_{obs} will increase if there are large discrepancies between the forecast based on the model, $\bar{\mathbf{U}}$ and the observed data, $\mathbf{u}_{\text{obs}}$. Thus D^2_{obs} can be referred to the theoretical values of the χ^2 distribution with nL degrees of freedom. Note also that D^2_{obs} is the Mahalanobis distance when the distributions of $\mathbf{U}_{\text{obs}}$ and $\bar{\mathbf{U}}$ have the common covariance matrix $\hat{\Sigma}$.

References

Cressie, N., Kaiser, M. S., Daniels, M. J., Aldworth, J., Lee, J., Lahiri, S. N., Cox, L. (1999). Spatial Analysis of Particulate Matter in an Urban Environment. In *GeoEnv II: Geostatistics for Environmental Applications*, (Eds. J. Gmez-Hernndez, A. Soares, R. Froidevaux)) Kluwer:Dordrecht, pp 41–52.

Kibria, B., Golam, M. Sun, L., Zidek, J. V., Le, N.D. (2002). Bayesian spatial prediction of random space-time fields with application to mapping $PM_{2.5}$ exposure. *Journal of the American Statistical Association*, **97**, 112–124.

Le, N. D., Sun, W., Zidek, J. V. (1997). Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society*, Series B, **59**, 501–510.

Mardia K.V., Goodall C., Redfern E.J., and Alonso F.J. (1998) The Kriged Kalman filter (with discussion). *Test*, **7**, 217–252.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2004) Spatio-temporal modeling of fine particulate matter. Available from `www.maths.soton.ac.uk/staff/Sahu`.

Sahu, S. K. and Mardia, K. V. (2005). A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society*, Series C, 223–244.

Smith, R. L., Kolenikov, S. and Cox, L. H. (2003). Spatio-Temporal modelling of $PM_{2.5}$ data with missing values. *Journal of Geophysical Research-Atmospheres*, **108** D24 9004, doi:10.1029/2002JD002914.

Zidek, J. V., Sun, L., Le, N., Zkaynak, H.(2002). Contending with space-time interaction in the spatial prediction of pollution: Vancouver's hourly ambient PM_{10} field. *Environmetrics*, **13**, 595–613.